Pima Indian Diabetes Data Analysis

By: Mildred Monsivais
July 30 - September 25

Table of Contents

Introduction

In this project I centered on applying exploratory data analysis to the relationship of which variables can predict the possibility of getting diabetes. It was based on exploring whether certain diagnostic measurements predict if the patient has diabetes in women at the age of 21year and older of the Pima Indian heritage. The dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. I analyzed the Pima Indian Women data set in python and imported the libraries matplotlib, seaborn, numpy and pandas to help my analysis. The libraries helped me plot a variety of different graphs in python which include a heat graph, histogram, pie chart, a regression graph, box plot and a bar graph. The visualization helped explored what the effect of one variable differ from another variable. By using graphs to visualize the different measured variables, it helps me interpret and gain insight into what variables are affected when having diabetes comparing to a non-diabetic individual. This project was analyzed in python because I had previous knowledge on how to use this language. In addition, python is also a flexible language that can be used for data manipulation. By using graphs to interpret large data sets is an external aid to improve interpretation on diabetic individuals and potential gain insight to a real-life problem.

Goal

The goal of the project is to gain maximum insight into the data set and understand the connection of the measured variables in cohorts with diabetes. With a large data set I was encouraged to explore all the data provided. This allowed me to critically think of a process when performing an investigation to discover any patterns between diabetes and the measured variables. Using visualizations to better understand diabetes and find clues about the

tendencies of the data is very important to better understand this health issue. In the short term this project challenged me to learn and practice using different libraries which I had never previously used before. In addition, helping me practice and improve my programming skills. Analyze data in the long run will help me work in a collaborative research setting and explore data collection in microorganisms such as bacterium, virus, and fungus to see how it can be beneficial to human health.

Data

This project consists of exploratory data analysis on the Pima Group Heritage, that several multiple predictor factors that affect the possibility of getting type 2 diabetes. The data set was collected from the National Institute of Diabetes and Digestive and Kidney Disease. The information was collected from female patients older than 21 years and specifically from the Pima Indian heritage. The Pima Indian heritage is a group of Native Americans that are living in Arizona. In this area there was a shift of traditional agriculture crops to processed foods and a decline of physical activity. For this reason, data was collected from this subject group to analyze the prevalence of getting type 2 diabetes. The whole data set consist of 768 instances(individuals) and 8 attributes. The 8 attributes include pregnancies, glucose, blood pressure, skin thickness, insulin, BMI, diabetes pedigree function, age and outcome.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

Figure 1: The first 5 rows of the Pima Group Heritage dataset. There is 8 measured variables which include pregnancies, glucose, blood pressure, skinthickness, insulin, BMI, diabetes pedigree function and age.

Pregnancies was measured by the number of times the women had a pregnancy. Glucose was measured by the plasma glucose concentration in a 2 hour oral glucose test. Blood pressure was measured by the diastolic blood pressure in millimeter of mercury. Skin thickness is measured by the triceps skin fold thickness in millimeters. BMI is measured by the body mass index which is the weight in kilograms divided by height in meters squared. The diabetes pedigree function is a score in the likelihood of diabetes based on family history. Age is the length of time alive measured at the beginning of birth. All these measured factors can predict whether a person is more likelihood is getting diabetes.

Means

The mean is used to average out the central tendency of the data. The overall data didn't contain any extreme outliners which was better suited for the measurement of the central tendency. I examined all of the attributes means to compare the variable and control group. In the data set outcome 1 means having diabetes and outcome 0 not having diabetes. Using the mean to interpret the predictor variables with the outcome of getting insight to which variable is being more affect is very important.

The average number pregnancies for people with diabetes is 4.85. This can be rounded up to a whole number of 5 to make the number of individuals more understandable. The control group had an average of 3.29 pregnancies as 3.29 can be rounded down to 3 pregnancies. This significance of having 2 more pregnancies on average for individuals with diabetes can enlighten questions of, does having more pregnancies cause a higher risk factor of

getting type 2 diabetes. The other attribute explored was the average glucose level for outcome 1 group was 141.25 and non-diabetics is 109.98. Generally speaking, people with diabetes have higher glucose level than individuals without. The next variable explored was the average BMI for outcome 1 which was 35.14 compared to the BMI for the outcome of 0 was 30.30. A characteristic of type 2 diabetes is having high insulin and insulin is fat-storage hormone which can account for an larger average value of BMI than non-diabetics. I wanted to examine categorical ages of a person with an outcome of 1 in contrast with an outcome of 0. The average age of a person with an outcome of 1 is 35.1 which can be rounded down to 35 years old. To give insight about which age on average as higher outcome rate of 1, I explored how many individuals in each age category are diabetic compared to non-diabetic. According to figure 2 on average the age group of 31 has a higher insulin level of 111.1 than age group 21 having a 73.6 insulin rate.

	Age	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction
0	21	1.079365	108.317460	65.936508	19.349206	73.634921	27.817460	0.433825
1	22	1.555556	108.208333	63.722222	20.486111	74.486111	29.509722	0.430625
2	23	1.578947	111.578947	64.315789	22.368421	118.026316	31.502632	0.438579
3	24	1.891304	117.891304	64.956522	25.934783	88.021739	32.569565	0.393565
4	25	1.770833	110.083333	59.666667	23.958333	82.895833	31.943750	0.600500
5	26	1.969697	118.212121	64.181818	23.666667	90.878788	34.915152	0.413455
6	27	2.562500	115.281250	73.500000	18.375000	63.125000	31.950000	0.471750
7	28	3.028571	119.914286	68.314286	23.628571	94.600000	33.642857	0.459629
8	29	3.310345	127.379310	68.241379	21.000000	88.793103	33.541379	0.408897
9	30	3.619048	122.285714	64.857143	18.904762	82.666667	30.033333	0.367238
10	31	3.875000	126.958333	64.375000	20.000000	111.166667	34.016667	0.589583

Figure 2: Averages of the 8 measured variables from ages 21 to 31.

Comparing different predicted variables can bring insightful question if whether one variable is related to predicting a person likelihood of getting type 2 diabetes.

Correlations

Correlations are a statistical technique that shows how strongly pairs of variables are related to one another. I decided to use correlation to explore predicted variables with one another. A correlation heat map uses colored cells in a monochromatic scale to show two discrete attributes in the Pima Group Diabetes dataset. It was more transparent and clearer to use one graphical representation to represent all the variables paired with one another. The first 8 variables studied in the pima data where being explored to predict the tendency of a person to develop type 2 diabetes.

In an ideal world there are perfect positive correlation meaning that a correlation coefficient is 1. This implies that as one variable increases the other variable increases as well. In contrast a negative correlation means that the two attributes move in opposite directions, as one variable increases the other variable decreases. A zero correlation suggest that the two attributes have no connection at all. According to the heatmap the highest correlation coefficient of pregnancy and age was 0.54 implying the older you get the more pregnancies an individual has.

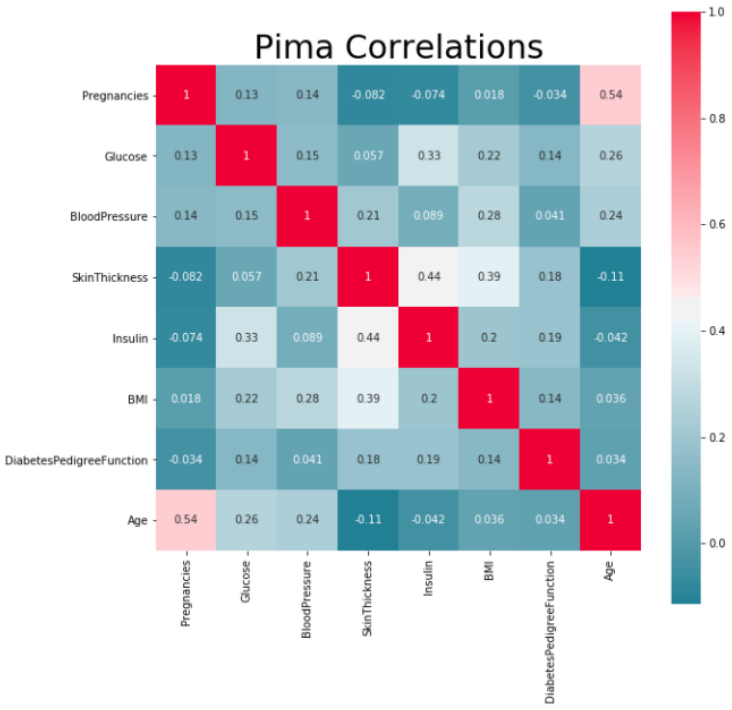

Figure 3: Heatmap correlation using all 8 attributes correlated by each variable. Correlation is indicated by the colored cells indicated by the monochromatic scale.

This is apprehensive as the longer lived the more likelihood the chances of having more children. A second correlation that was considered is to be fairly strong in the heatmap was between BMI and the thickness of the skin. The correlation of 0.39 can be interpreted of how an increase in BMI can causes a higher resistance of skin thickness.

Visualization

Visuals and graphs help me understand and remember key points of the project. I wanted to understand the key distribution among people that are predicted to having type 2 diabetes as to non-diabetics. Explaining the pima data set helped me find patterns and relationships in the analysis.

A simple way to express the relationships between two variables within the entire data set was by using a pair plot. A pair plot is a way to visualize relationships between each variable

containing pregnancies, glucose, blood pressure, skin thickness, insulin, BMI age, outcome and diabetes pedigree function.

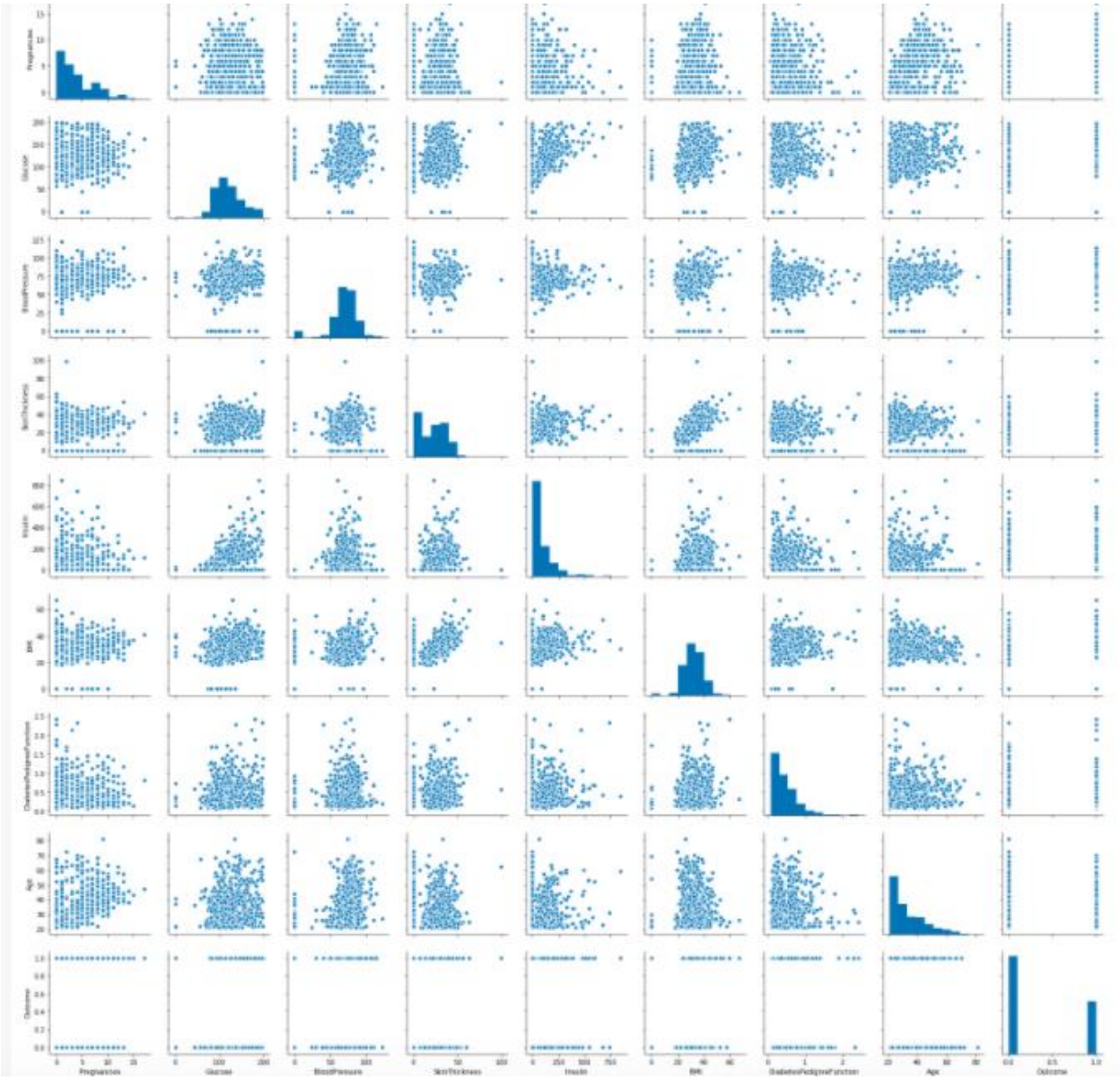

Figure 4: A pair plot using the 8 measured variables from the dataset plotted against each variable. The two graphs included in a pair plot are histograms and scatter plots.

It's an instant examination of each of the data relationship between two variables. There are two types of graphs a pair plot shows which are histograms and scatter plots. Histograms are used for numerical features to display the distribution of the values, as scatter plots are

used for numerical values. In this analysis the pair plot shows a positive correlation between skin thickness and BMI. This indicate how an increase in BMI can overall affect skin thickness.

A pie chart was used to show the percentage of diabetic and non-diabetic in the Pima Heritage dataset. The overall results of people with and without diabetes is presented at the beginning of the analysis which is done to inform the audience about an overall finding before analyzing the different variables. The finding shows that 69.5% of the individuals in the dataset are diabetic and 30.5% are non-diabetic.

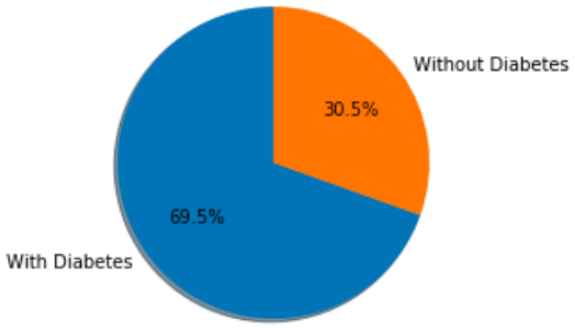

Figure 5: Pie chart shows the percentage of diabetics and non-diabetics in the Pima dataset.

The data is biased towards diabetic's women which was something to consider before analyzing the predicted variables.

Regression lines are used to show the relationship between two variables in a linear equation. The regression line is consisted to be the best fitting line in the data set. The diabetic count was the dependent variable because diabetic individuals were the ones being recorded. The two independent variables shown in the graphs were age and blood pressure. The first relationship I wanted to see more in depth was how age affects diabetes. Age is the independent variable which is plotted on the x axis as the diabetic count is dependent on the y

axis. On average the regression line is decreasing so as age progresses there contains a less diabetic count in the population.

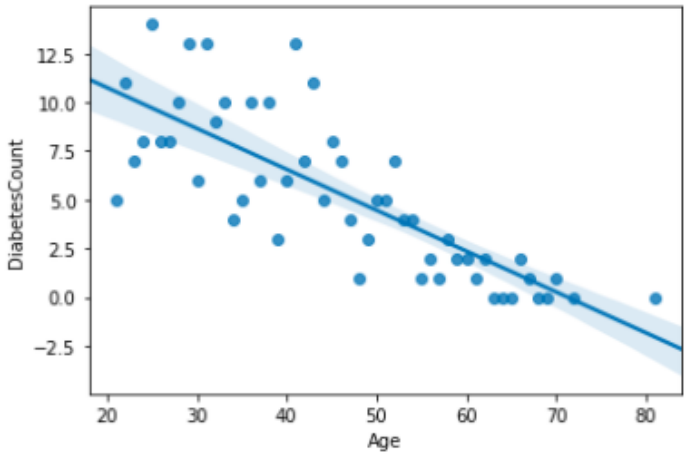

Figure 6: A scatter plot illustrating the relationship between the age causing a change in the offset of diabetes in the population. A downward trend showed by the best fit line indicating a diabetes consistently decreases throughout time.

This could be explained by the distribution of the age in the dataset. There is less data collected from women above 50 years old than women that are older than 21 which could have caused the overall trend for the regression line. The next regression graph can be associated with unpredictable results as I was certain that there was going to be an association between blood pressure and diabetes. Analyzing the two variables illustrated an inconclusive graph as the data points were too scattered between blood pressure and the number of individuals with diabetes to be inconclusive.

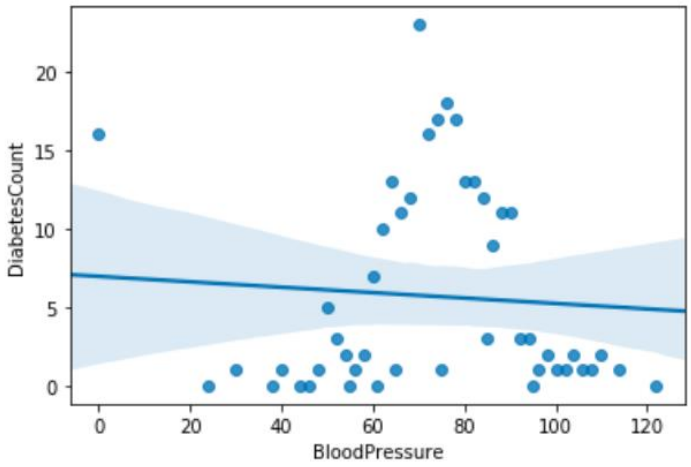

Figure 6: A scatter plot representing the blood pressure and diabetes count with a best fit regression line to assess the relationship between the two variables.

The next two variables I wanted to explore was centered on the individuals having diabetes and the proportion of number of pregnancies based on the age. The box plot shows that on average people with diabetes have higher age distributed for each category of pregnancies than non-diabetics.

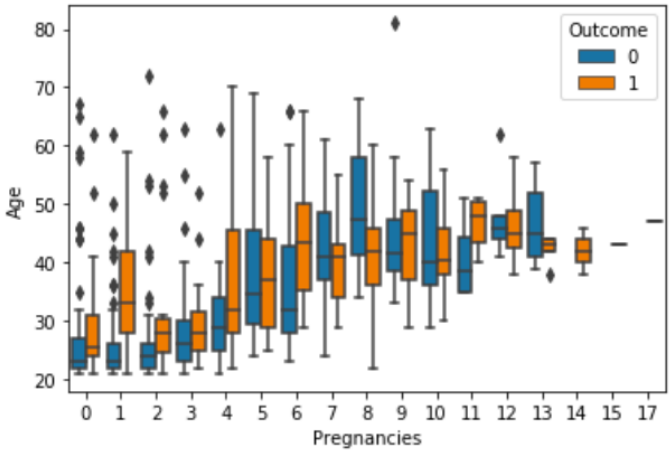

Figure 7: A box plot displaying the summary of diabetic and non-diabetic categorized by pregnancies. Outcome 1 indicates diabetics as outcome 0 are non-diabetics. The separation of the maximum and minimum is dependent on the age of outcome 1 and outcome 0 categorized by pregnancies.

This could account for the fact that having more pregnancies at an older age in the long term can affect the health of an individual.

Conclusion

This projected helped me analyze diabetic and non-diabetic data for the Pima Heritage population. I used visuals to summarize the main characteristics of the measured variables within each group. By using matplotlib, seaborn, numpy and pandas I found that there was a strong correlation between pregnancies and diabetes. In addition, there was a strong correlation between skin thickness and BMI which on average diabetic women had a higher BMI. Using different visuals helped me assess connections between the measure's variables and diabetic individuals. Furthermore, practicing how to use different libraries in python can help me be prepared for my future career goals in analyzing data sets in order to uncover patterns to understand the issue more thoroughly. In the future I want to take the supplements I learned from this project and build a machine learning model to predict with the measured variables whether or not patients in the dataset could have diabetes or not.